Application and evaluation of a DENCLUE clustering algorithm for landslide susceptibility mapping

Jingguo Dai[1], Deborah Simon Mwakapesa[2*]
(1. *School of Information Engineering, Shaoguan University, Shaoguan, Guangdong 512005, China*;
2. *School of Civil and Surveying Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi 341000, China*)

Abstract: In this research, a clustering algorithm named the DENsity-based CLUstEring (DENCLUE) algorithm is applied and evaluated for landslide susceptibility mapping in Baota District, China. The proposed methodology works well with large datasets, can handle noise effectively, and can obtain clusters of different types. A dataset containing landslide records and 7 landslide influencing factors was prepared for modeling. 453 well-scattered clusters of various shapes and sizes were obtained from clustering the study area mapping units using the DENCLUE algorithm. The natural breaks method was adopted to classify the clusters into 5 susceptibility classes using landslide density, eigenvalues, and geology expertise. A map was then constructed showing landslide susceptibility in the area, which presented a significant assessment of landslide susceptibility in the Baota District. Moreover, the model was evaluated and compared with DBSCAN, K-means, and KPSO clustering algorithms based on statistic metrics. The results indicated that DENCLUE obtained the highest performance, and was thus, the best among others. The constructed map can serve as a tool to identify safety areas within the Baota district, which are suitable for habitation and economic activities.

Keywords: Landslide, Landslide susceptibility mapping, Clustering, DENCLUE algorithm, Baota
DOI: 10.33440/j.ijpaa.20220501.191

Citation: Dai J G, Mwakapesa D S. Application and evaluation of a DENCLUE clustering algorithm for landslide susceptibility mapping. Int J Precis Agric Aviat, 2022; 5(1): 35–40.

1 Introduction

Landslide is a frequent and common natural catastrophe that results in lives, and social, environmental, and economic losses annually (Figure 1). Mapping the regions vulnerable to landslides is the preliminary phase for landslide and its risk assessments. Landslide susceptibility is the possibility of a landslide to occur in a certain region under terrain conditions that assess the degree to which terrain can be susceptible to future landslides[1, 2].

Figure 1 Landslide events

During the last several years, a number of methods have been explored in constructing susceptibility maps[3-18]. Generally, these methods can be classified into supervised and unsupervised learning approaches[19]. The posterior supervised learning approaches such as classification algorithms can investigate the association between landslide susceptibility and landslide influencing factors based on landslide inventory data, i.e., labeled training data[20-23]. However, those methods cannot work if there is not enough inventory data.

To make use of the limited data, clustering methods are proposed. Cluster analysis is an unsupervised learning approach of multivariate statistical data that is widely used in various fields including pattern recognition, social network analysis, medical studies, and image processing. Compared with the posterior methods, the unsupervised learning[19] approaches are more advantageous, as they can divide a multivariate dataset into distinct classes (called clusters) without using labeled training data, and can work with the available limited data.

For these reasons, this study presented a comprehensive application and evaluation of a clustering method named DENCLUE[24] in mapping the susceptibility of landslides in the Baota District. The method not only does not require training data but also, can work well with large datasets and can handle noise effectively. Nevertheless, it is a tentative but useful method for the present purpose when there is not enough information on landslide distribution.

2 Study area and Materials

2.1 Study Area

Baota District is found in the northern area of the Shaanxi Province in China. Its area is approximately 3556 km^2. The average annual temperature and average annual rainfall of are 7°C and 550 mm respectively. About 60% of the rains (heavy and continuous rains) occur between July and September, resulting in extreme soil erosion. Also, these seasons are recorded to have frequent landslide events[25, 26]. Also, the northern side of the area

Received date: 2022-11-27 **Accepted date:** 2022-12-23
Biography: Jingguo Dai, Professor, research interest: intelligent decision-making and machine learning, Email: 646257139@qq.com
***Corresponding author: Deborah Simon Mwakapesa**, Doctoral student, research interests: Machine learning, clustering algorithms, landslide susceptibility mapping, GIS. Email: mymlyc@163.com.

has dense vegetation while the southern side is sparsely vegetated. Generally, the climate of the area and its terrain condition support the frequent occurrence of landslides. Figure 2 shows the study area and some recorded landslide sites.

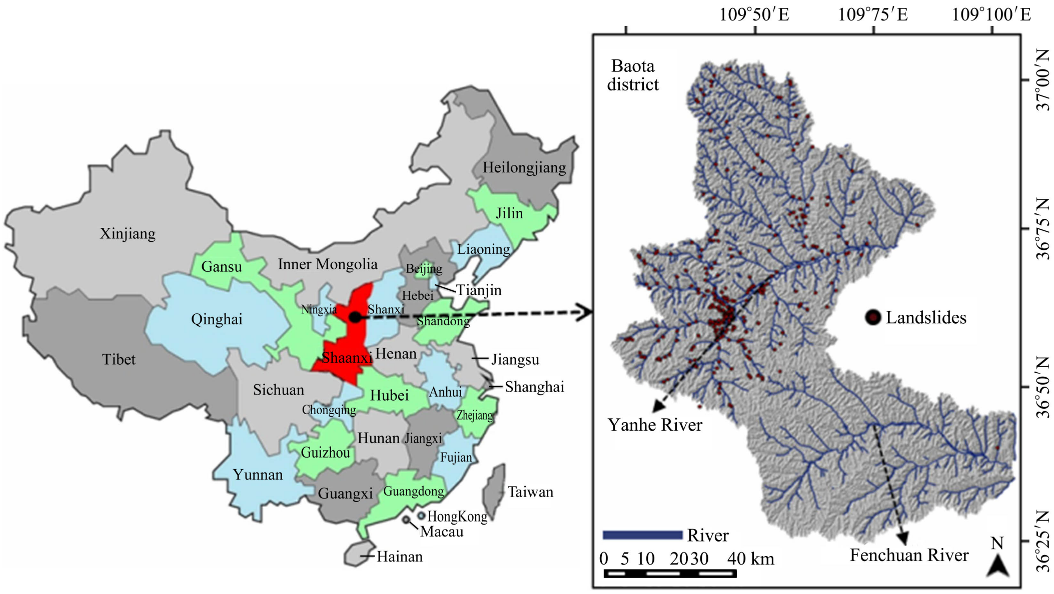


Figure 2 The study area and recorded landslide sites

2.2 Landslide Database

The landslide database collected from the Xi'an Center of Geology Survey of the Baota District was used to examine landslide susceptibility in this research. The database was obtained based on the analysis of 1081 locations using remote sensing images, and sites survey data. From this, 293 landslide events (their locations are spotted in Figure 2) were recorded which were used for model evaluation. Also, from the study area, a sample of 213 non-landslide locations was selected randomly for model evaluation.

2.3 Landslide Influencing Factors (LIF)

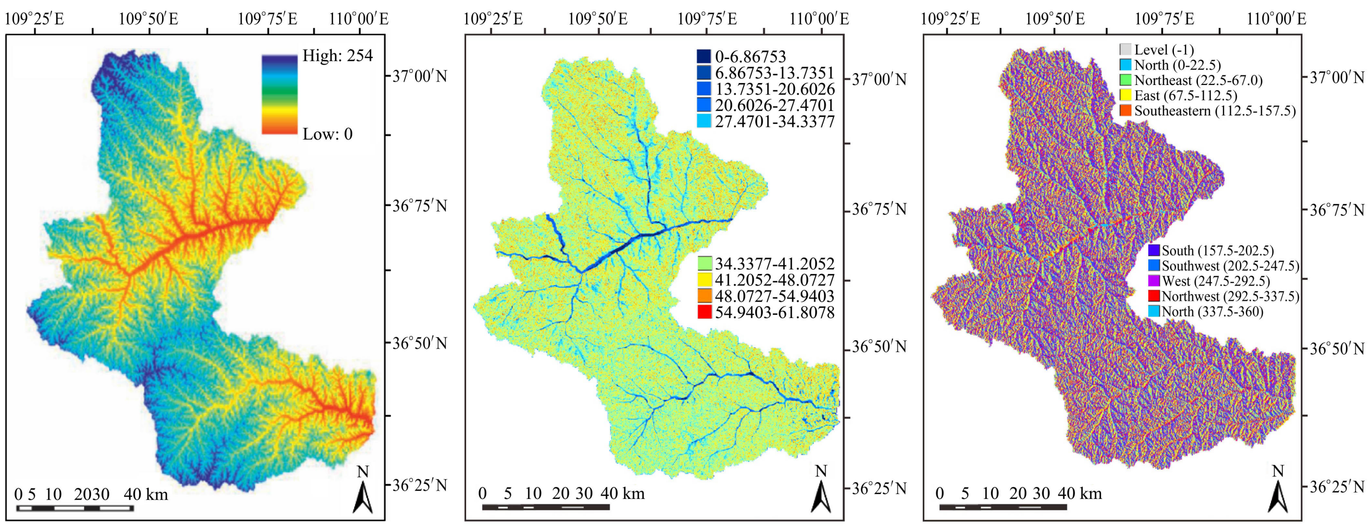
The landslide susceptibility mapping includes various factors that are associated with landslide occurrences. Hence, decide on appropriate factors is very vital in LSM modeling. To construct the LSM model, this study selected 7 LIF (that were of 4 types) based on previous studies that were developed and conducted in the same region. These factors are geomorphologic and topographic factors (elevation, slope angle, slope aspect), geology factor (lithology), underlying surface (NVDI), and activating factors (rainfall). Those studies that used these factors showed that the factors have a strong relationship with the recorded landslides, and thus they represent well the study area[26-28]. Data for these factors were extracted from DEM (which was obtained from topographic maps) at a scale of 1:10000 and 25 m resolution (Figure 3a–g).

The applied dataset was analyzed in ArcGIS 10.2 software. The whole region was partitioned into 5,672,922 mapping units of both landslide and non-landslide sites, each comprising the 7 layers of LIF.

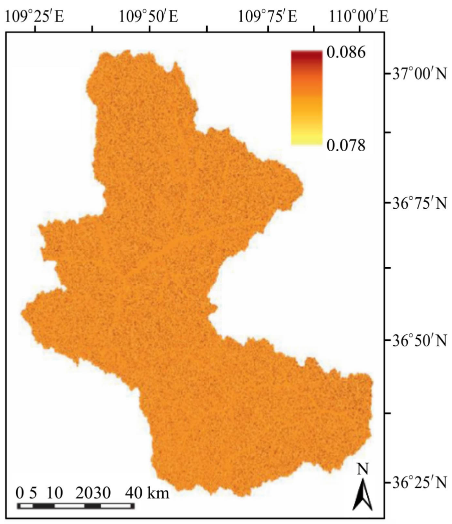
3 Research Methods

3.1 DENCLUE Algorithm

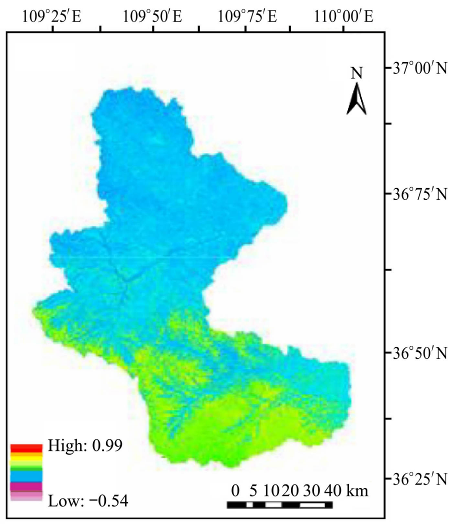
DENCLUE[24] is an algorithm that relies on a group of density distribution functions. The influence of every point is formally demonstrated using a function, namely influence function, which defines the influence of a point within its region. The density of the data area can be demonstrated as the summation of the influence functions used at some points. Clusters are then obtained numerically by finding density attractors, which are the local maxima of the overall density function. It partitions the n-dimensional dataset into small, adjacent, non-overlapping



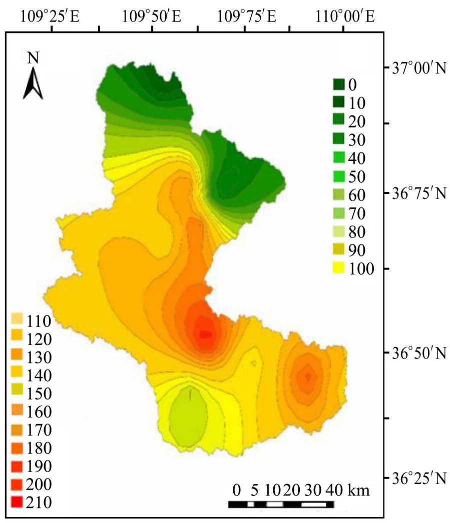
a. Elevation b. Slope angle c. Slope aspect



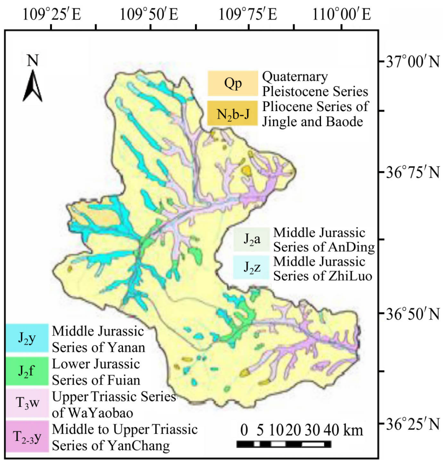
d. Profile curvature



e. NDVI



f. Rainfall



g. Lithology

Figure 3 Continued

n-dimensional hyper-cubes, but only where there are points. The identification of clusters then continues in the hyper-cubes that are sufficiently dense in points. DENCLUE is good for clustering large datasets containing a lot of noise and can define clusters of different types. The framework of the algorithm is described below.

Let u and v point in f^n, an n-dimensional input data. DENCLUE takes parameters: ξ (a threshold; when the density function at point u exceeds this threshold, u is density attracted to a local maximum u^* of the density function, which is called a density attractor of u), and σ (the standard deviation of a Gaussian that defines the influence of a point in its neighborhood).

Definition 1: The influence function of a point v on u is a function; $f_B^v : f^n \to R_0^+$ that is defined using a basic influence function f_B:

$$f_B^v(u) = f_B(u, v)$$

f_B is a Gaussian influence function (GIF) f_{GIF} that is obtained based on the Euclidean distance $d(u, v)$ between two points in a neighborhood, defined as:

$$f_{GIF}(u, v) = e^{-(\frac{d(u,v)^2}{2\sigma^2})}$$

Definition 2: The density function f_{GIF}^D at a point $u \in f^n$, is the sum of influence functions of all points. It is given by the equations below:

$$f_{GIF}^D(u) = \sum_{i=1}^{n} e^{-(\frac{d(u,v)^2}{2\sigma^2})}$$

Definition 3: The attracted point is a point that forms a path with the density attractor.

Principally, this algorithm works in two steps, the pre-clustering step, and the clustering step.

Step 1: *Pre-processing*

1. Take a dataset and construct an n-dimension grand hypercube that contains all the points.

2. Assign the points to a smaller n-dimensional hypercube (of edge length 2σ) which divides the grand hypercube.

3. Count the number of points in a hypercube and identify a highly populated hypercube. If the number is more than ξ, the hypercube is considered highly populated (HPH).

4. Calculate the mean of each populated hypercube.

5. Use the distance between their means to identify the connection between each HPH, and other hypercubes (highly or just populated hypercube). If $d(mean_{c_1}, mean_{c_2}) < 4\sigma$, then the two hypercubes are connected.

Step 2: *Clustering*

1. Find the density attractors by scanning the HPH. If a point in HPH is not yet density attracted to an attractor, then its attractor is computed. The found density attractor is then registered.

2. Form a cluster for each registered density attractor, and add each point to its corresponding cluster, since the density attractor of each point is now known, and the attractor is associated with a cluster.

3. Connect density attractors that have the same path to create

the final clusters.

Note: The term "points" in the algorithm refers to "mapping units" in LSM modeling.

During algorithm execution, the values of the mapping units were first standardized to range between –1 and 1. Then, these standardized values were then used as inputs into the DENCLUE algorithm.

3.2 Determination of Landslide Susceptibility Classes

In this study, the natural breaks technique (NBT)[1] in the ArcGIS software is applied to classify the clusters obtained from the DENCLUE algorithm into 5 classes (very high, high, moderate, low, and very low) using landslide density (L). L is found by computing the number of landslides/km^2 of a mapping unit in a cluster[29]. To do this, using the sorting tool in the ArcGIS software, the mapping units and landslides in each cluster are sorted and counted, followed by calculating the L in each cluster. Then the assumption that when landslide density is high, the susceptibility is also high is applied to clarify the susceptibility classes among the clusters. Moreover, for the clusters with a landslide density value of zero, the susceptibility classes are determined based on geology expertise on eigenvalues which describes the characteristics of the mapping units in the clusters.

3.3 Model's Performance Assessment

To assess the overall performance of the model the accuracy, specificity, and sensitivity were applied. These metrics are computed based on four indices: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP is the true predicted landslide value; TN is the true predicted non-landslide value; FP is the landslide value that is incorrectly predicted; and FN is the incorrectly predicted non-landslide value[30-32].

$$Accuracy = (TP + TN) / (TN + TP + FP + FN)$$

$$Sensitivity = TP / (TP + FN)$$

$$Specificity = TN / (TN + FP)$$

Furthermore, K-means[33], DBSCAN[28], and KPSO[34] clustering algorithms are applied for comparison based on data from Baota District.

4 Results and Discussion

4.1 DENCLUE Clustering Analysis

By following the procedure described above, the algorithm generated 453 well-scattered clusters of varying shapes and sizes. Each cluster contained mapping units that had similar geology and geomorphology characteristics that were different from the mapping units of other clusters. These showed that the DENCLUE algorithm has good clustering properties.

4.2 Construction of Landslide Susceptibility Map

The eigenvalues, landslide density, and susceptibility classes of some clusters are shown in Table 1. Among the 453 clusters, 19%, 21%, 26%, 18%, and 16% of the clusters were in moderate, high, very high, low, and very low susceptibility classes respectively. The results are illustrated in Figure 4. As it can be noted in the figure, very highly and highly susceptible clusters dominated the upper part of the region while low and very low susceptible clusters were in the lower part of the region. These results are in close agreement with the landslide distribution shown in the area map, indicating that the algorithm has good prediction capability.

Table 1 Landslide density and Susceptibility classes of some clusters

Cluster No.	Attribute Values							Landslide Density			Landslide susceptibility class
	Elevation	Slope angle	Profile curvature	Slope aspect	Lithology	NDVI	Rainfall	Area /km^2	Number of landslides	L /km^2	
1	35.21	21.87	0.024	S	III	0.63	24–231	9.52	1	0.1	Very Low
2	21.35	26.19	0.023	SE	I	0.52	25–188	5.53	5	0.9	High
…	…	…	…	…	…	…	…	…	…	…	…
256	19.97	38.23	0.57	N	II	0.54	22–185	13.25	0	0	Identified by expert
…	…	…	…	…	…	…	…	…	…	…	…

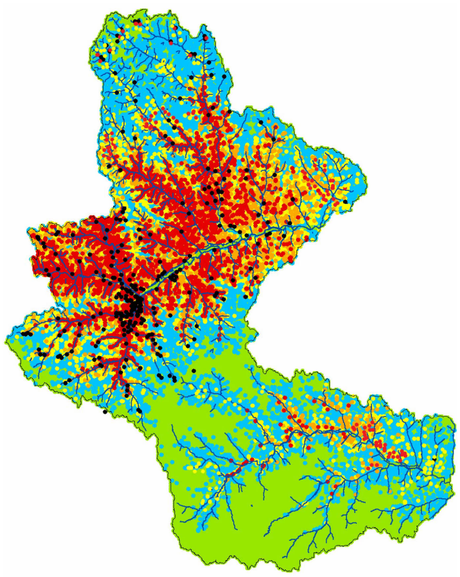


Figure 4 Landslide susceptibility Map based on DENCLUE algorithm (Red: very high; yellow: high; Orange: moderate; Blue: low; Green: very low)

4.3 The Assessment Results

The model constructed in this research was evaluated based on 293 landslide sites and 213 non-landslide sites (a total of 506). The evaluation is based on sensitivity, specificity, and accuracy, (shown in Table 2). Results analysis showed that the DENCLUE performed well with a Sensitivity= 87.37% and Specificity = 87.32%. The high performance of the model was also shown by its high prediction accuracy value of 87.35%. These results were compared with those of the K-means, DBSCAN, and KPSO models. The DENCLUE model got the best results, followed by DBSCAN (Sensitivity = 0.8669, Specificity = 0.8638, Accuracy = 0.8656), and KPSO (Sensitivity = 0.6724, Specificity = 0.6479, Accuracy = 0.6621). K-means had the lowest performance with Sensitivity, Specificity, and Accuracy of 0.6246, 0.6761, and 0.6466, respectively. These results indicated that DENCLUE good prediction capability in landslide susceptibility mapping. This high performance was greatly supported by the advantageous capability of the DENCLUE model to work well with large and high dimensional datasets, the ability to deal with noise successfully, and its ability to define clusters of different types, that is, center-defined clusters and arbitrary-shaped clusters.

Table 2 Assessment Results

Models	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy
DENCLUE	256	186	27	37	0.8737	0.8732	0.8735
DBSCAN	254	184	29	39	0.8669	0.8638	0.8656
KPSO	197	138	75	96	0.6724	0.6479	0.6621
K-means	185	130	83	108	0.6314	0.6103	0.6225

5 Conclusion

This study explored and evaluated the performance of a clustering algorithm named DENCLUE in LSM modeling for Baota District, China. The algorithm was executed based on a dataset that included landslide inventory and 7 landslide influencing factors. Also, during evaluation and comparison, the proposed model acquired the best performance with sensitivity, specificity, and accuracy values of 87.37%, 87.32, and 87.35%, respectively. The proposed method can work well with a large dataset, can handle noise effectively, and can identify clusters of different types, features that can rarely be found in other clustering algorithms. Moreover, the developed map can aid and contribute to the landslide assessment and management strategies to ensure proper land use planning and a safe environment.

[References]

[1] Mersha T, Meten M. GIS-based landslide susceptibility mapping and assessment using bivariate statistical methods in Simada area, northwestern Ethiopia. Geoenvironmental disasters, 2020; 7(1): 1–22. doi: 10.1186/s40677-020-00155-x.

[2] Akinci H, Zeybek M. Comparing classical statistic and machine learning models in landslide susceptibility mapping in Ardanuc (Artvin), Turkey. Natural Hazards, 2021; 108(2): 1515–43. doi: 10.1007/s11069-021-04743-4.

[3] Pourghasemi H. R., Teimoori Yansari Z., Panagos P, Pradhan, B. Analysis and evaluation of landslide susceptibility: a review on articles published during 2005–2016 (periods of 2005–2012 and 2013–2016). Arabian Journal of Geosciences, 2018; 11(9): 1–12. doi: 10.1007/s12517-018-3531-5.

[4] Broeckx J, Vanmaercke M, Duchateau R, Poesen J. A data-based landslide susceptibility map of Africa. Earth-Science Reviews, 2018; 185: 102–21. doi: 10.1016/j.earscirev.2018.05.002.

[5] Dias H C, Hölbling, D, Grohmann C H. Landslide susceptibility mapping in Brazil: a review. Geosciences, 2021; 11(10): 425. doi: 10.3390/geosciences11100425.

[6] Ma Z, Mei G, Piccialli F. Machine learning for landslides prevention: a survey. Neural Computing and Applications, 2021; 33(17): 10881–907. doi: 10.1007/s00521-020-05529-8.

[7] Bragagnolo L, da Silva R, Grzybowski J. Artificial neural network ensembles applied to the mapping of landslide susceptibility. Catena, 2020; 184: 104240. doi: 10.1016/j.catena.2019.104240.

[8] Ganga A, Elia M, D'Ambrosio E, Tripaldi S, Capra G F, Gentile F, Sanesi G. Assessing Landslide Susceptibility by Coupling Spatial Data Analysis and Logistic Model. Sustainability, 2022; 14(14): 8426. doi: 10.3390/su14148426.

[9] Zhang K, Wu X, Niu R, Yang K, Zhao L. The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area, China. Environmental Earth Sciences, 2017; 76(11): 1–20. doi: 10.1007/s12665-017-6731-5.

[10] Ngandam Mfondoum A H, Wokwenmendam Nguet P, Mefire Mfondoum J V, Tchindjang M, Hakdaoui S, Cooper R, et al. Adapting sudden landslide identification product (SLIP) and detecting real-time increased precipitation (DRIP) algorithms to map rainfall-triggered landslides in Western Cameroon highlands (Central-Africa). Geoenvironmental Disasters, 2021; 8(1): 1–26. doi: 10.1186/s40677-021-00189-9.

[11] Laimer H J. Anthropogenically induced landslides–A challenge for railway infrastructure in mountainous regions. Engineering Geology, 2017; 222: 92–101. doi: 10.1016/j.enggeo.2017.03.015.

[12] Solanki A, Gupta V, Joshi M. Application of machine learning algorithms in landslide susceptibility mapping, Kali valley, Kumaun Himalaya, India. Geocarto International, 2022(just-accepted): 1–27. doi: 10.1080/10106049.2022.2120546.

[13] Dou J, Yunus A P, Merghadi A, Shirzadi A, Nguyen H, Hussain Y, et al. Different sampling strategies for predicting landslide susceptibilities are deemed less consequential with deep learning. Science of the total environment, 2020; 720: 137320. doi: 10.1016/j.scitotenv.2020.137320.

[14] Pham B T, Prakash I. Evaluation and comparison of LogitBoost Ensemble, Fisher's Linear Discriminant Analysis, logistic regression and support vector machines methods for landslide susceptibility mapping. Geocarto International, 2019; 34(3): 316–33. doi: 10.1080/10106049.2017.1404141.

[15] Nohani E, Moharrami M, Sharafi S, Khosravi K, Pradhan B, Pham B T, et al. Landslide susceptibility mapping using different GIS-based bivariate models. Water, 2019; 11(7): 1402. doi: 10.3390/w11071402.

[16] Pham B T, Prakash I, Dou J, Singh S K, Trinh P T, Tran H T, et al. A novel hybrid approach of landslide susceptibility modelling using rotation forest ensemble and different base classifiers. Geocarto International, 2020; 35(12): 1267-92. doi: 10.1080/10106049.2018.1559885.

[17] Nanehkaran Y A, Mao Y, Azarafza M, Kockar M K, Zhu H H. Fuzzy-based multiple decision method for landslide susceptibility and hazard assessment: A case study of Tabriz, Iran. Geomechanics and Engineering, 2021; 24(5): 407–18. doi: 10.12989/gae.2021.24.5.407.

[18] Mao Y.-m., Zhang, M.-s., Wang, G.-l. and Sun, P.-p. Landslide hazards mapping using uncertain Naïve Bayesian classification method. Journal of Central South University, 2015; 22(9): 3512–20. doi: 10.1007/S11771-015-2891-1.

[19] Su C, Wang B, Lv Y, Zhang M, Peng D, Bate B, et al. Improved landslide susceptibility mapping using unsupervised and supervised collaborative machine learning models. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 2022; 1–19. doi: 10.1080/17499518.2022.2088802.

[20] Youssef A M, Pourghasemi H R, Pourtaghi Z S, Al-Katheeri M M. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin. Asir Region, Saudi Arabia. Landslides, 2016; 13(5): 839–56. doi: 10.1007/s10346-015-0614-1.

[21] Park S J, Lee C W, Lee S, Lee M J. Landslide susceptibility mapping and comparison using decision tree models: A Case Study of Jumunjin Area, Korea. Remote Sensing, 2018; 10(10): 1545. doi: 10.3390/rs10101545.

[22] Sheng M, Zhou J, Chen X, Teng Y, Hong A, Liu G. Landslide Susceptibility Prediction Based on Frequency Ratio Method and C5. 0 Decision Tree Model. Frontiers in Earth Science, 2022; 889. doi: 10.3389/feart.2022.918386.

[23] Rong G, Alu S, Li K, Su Y, Zhang J, Zhang Y, et al. Rainfall induced landslide susceptibility mapping based on bayesian optimized random forest and gradient boosting decision tree models—a case study of shuicheng county. China. Water, 2020; 12(11): 3066. doi: 10.3390/w12113066.

[24] Hinneburg A, Keim D A. An efficient approach to clustering in large multimedia databases with noise. Bibliothek der Universität Konstanz, 1998.

[25] Guo L, Di L, Li G, Luo Q, Gao M. GIS-based detection of land use transformation in the Loess Plateau: A case study in Baota District, Shaanxi Province, China. Journal of Geographical Sciences, 2015; 25(12): 1467–78. doi: 10.1007/s11442-015-1246-z.

[26] Zhang M.-s, Liu J. Controlling factors of loess landslides in western China. Environmental Earth Sciences, 2010; 59(8): 1671–80. doi: 10.1007/s12665-009-0149-7.

[27] Hu J, Xu K, Wang G, Liu Y, Khan M A, Mao Y, et al. A novel landslide susceptibility mapping portrayed by OA-HD and K-medoids clustering algorithms. Bulletin of Engineering Geology and the Environment, 2021; 80(2): 765–79. doi: 10.1007/s10064-020-01863-2.

[28] Mao Y, Mwakapesa D S, Xu K, Lei C, Liu Y, Zhang M. Comparison of wave-cluster and DBSCAN algorithms for landslide susceptibility assessment. Environmental Earth Sciences, 2021; 80(22): 1–14. doi: 10.1007/s12665-021-09896-w.

[29] Yimin M, Yican L, Simon Mwakapesa D, Genglong W, Ahangari Nanehkaran Y, Asim Khan M, et al. Innovative Landslide Susceptibility Mapping Portrayed by CA-AQD and K-Means Clustering Algorithms. Advances in Civil Engineering, 2021; 2021. doi: 10.1155/2021/8846779.

[30] Ling C X, Huang J, Zhang H. AUC: a statistically consistent and more

discriminating measure than accuracy. City, 2003. doi: 10.5555/1630659.1630736.

[31] Ling S, Zhao S, Huang J, Zhang X. Landslide susceptibility assessment using statistical and machine learning techniques: A case study in the upper reaches of the Minjiang River, southwestern China. Frontiers in Earth Science, 2022; 10: 986172. doi: 10.3389/feart.2022.986172.

[32] Wang S, Lin X, Qi X, Li H, Yang J. Landslide susceptibility analysis based on a PSO-DBN prediction model in an earthquake-stricken area. Frontiers in Environmental Science, 2022: 1244. doi: 10.3389/fenvs.2022.912523.

[33] Wang Q, Wang Y, Niu R, Peng L. Integration of information theory, K-means cluster analysis and the logistic regression model for landslide susceptibility mapping in the Three Gorges Area, China. Remote Sensing, 2017; 9(9): 938. doi: 10.3390/rs9090938.

[34] Wan S, Yen J Y, Lin C Y, Chou T Y. Construction of knowledge-based spatial decision support system for landslide mapping using fuzzy clustering and KPSO analysis. Arabian Journal of Geosciences, 2015; 8(2): 1041–55. doi: 10.1007/s12517-013-1226-5.